CIS: Compound Importance Sampling Method for Protein-DNA Binding Site p-value Estimation

Y. Barash[1, 3], G. Elidan[1, 3], T. Kaplan[1, 2, 3] and N. Friedman[1]

[1]School of Computer Science & Engineering, The Hebrew University, Jerusalem 91904, Israel. [2]Hadassah Medical School, The Hebrew University, Jerusalem 91120, Israel. [3]These authors contributed equally to this manuscript.

ABSTRACT

Motivation: Transcription regulation involves binding of transcription factors to sequence-specific sites and controlling the expression of nearby genes. Given binding site models, one can scan the regulatory regions for putative binding sites and construct a genome-wide regulatory network. Several recent works demonstrated the importance of modeling dependencies between positions in the binding site. The challenge is to evaluate the statistical significance of binding sites using these models.

Results: We present a general, accurate and efficient method for this task, applicable to any probabilistic binding site and background models. We demonstrate the accuracy of the method on synthetic and real-life data.

Availability: The algorithm used to compute the statistical significance of putative binding sites scores is available online at http://compbio.cs.huji.ac.il/CIS/

Contact: *E-mail*: nir@cs.huji.ac.il

INTRODUCTION

Accurate detection of *cis*-regulatory elements in long DNA sequences is a central challenge in modern biology, as it offers a direct way for elucidating transcription regulation. Extensive efforts have been put in gathering known transcription factor binding sites, and in finding models that characterize them. These models facilitate systematic scans of genomic sequences to identify targets of transcription factors.

A fundamental challenge when performing such a genome-wide scan is to improve the model's accuracy in target prediction, and reduce the number of errors. This problem is further emphasized in eukaryotic genomes where binding sites appear in large intergenic regions. As a consequence, there is high probability of finding spurious binding sites due to the immense number of subsequences evaluated. To control the amount of this false positive noise in the predictions, we assign each possible site a score, and estimate its statistical significance, i.e. how likely it is to see a score that is at least as good by chance.

Formally, the p-value of a score S is the probability of achieving this score or higher according to the background distribution:

$$\boldsymbol{E}_{P_{BG}(X)}[1\{Score(X) \geq S\}]$$

where $1\{\}$ is the indicator function, $P_{BG}(X)$ is the background distribution over a random variable X that ranges over possible DNA motifs, and $Score$ is the score function. A common score is the log odds between the motif probability according to the binding site model, and its probability according to a background model.

A simple estimate of the p-value of a score S is by sampling from $P_{BG}(X)$, and then computing the fraction of samples that have score as high as S. Such a *naive sampling* procedure can reliably estimate p-values that are at least two orders of magnitude larger than the inverse of the number of samples. Thus, to estimate p-value of 10^{-3} we need about 10^5 samples. This is a problem when scanning long sequences where we need to compensate for multiple testing (Benjamini and Hochberg , 1995). Assuming a typical promoter length of size 500bp, these corrections result in the need for estimating p-values in the order of 10^{-5} or lower, rendering the naive sampling approach almost impractical, with at least 10^7 samples.

More sophisticated approaches either analytically derive the p-value (Wu *et al.*, 2000; Huang *et al.*, 2004), or use large deviation approximation (Bailey and Gribskov , 1998). These are effective for *probabilistic profile models* (also known as *PWMs*) that assume independence between nucleotides at different binding site positions.

Recently, several works demonstrated the importance of modeling transcription factor binding sites using richer probabilistic models that allow for inner-dependencies within the positions of a binding site (Barash *et al.*, 2003; King and Roth, 2003; Zhou and Liu, 2004). As these works show, by using such dependency models one can improve the accuracy of binding site identification. Yet, the question of assigning p-values for putative binding sites when using such dependency models remains open. Specifically, analytical methods that are designed for PWMs are not applicable for such richer models.

We present a general, accurate and efficient method for estimating p-values. Our *Compound Importance Sampling*

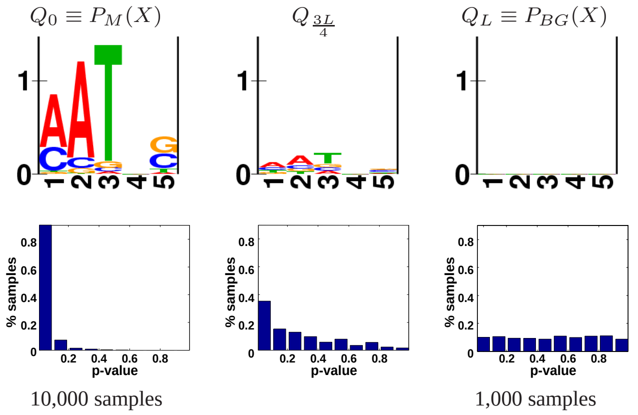

Fig. 1. Illustration of a proposal distribution $Q(X)$. The top panel shows the sequence logos ranging from the binding site model (left) to the background model (right), where the information content of each position is low. The bottom panel shows histograms of the p-value distributions of samples generated from each component.

(CIS) algorithm uses *importance sampling* (Gelman *et al.*, 1995), which allows us to conceptually mimic the naive sampling approximation using a significantly smaller number of samples. CIS does not require simplifying assumptions on the models for either the binding site or the background. We demonstrate the accuracy and efficiency of the CIS method on synthetic and real-life data, both for the case of simple position-independent models and for models that allow dependencies, where standard methods cannot be applied.

COMPOUND IMPORTANCE SAMPLING (CIS)

In the naive sampling approach we approximate the expectation $\boldsymbol{E}_{P_{BG}(X)}[1\{Score(X) \geq S\}]$, by computing the fraction of samples where $Score(X) \geq S$. Sampling directly from P_{BG} leads to a poor estimate of small p-values, since for most samples $1\{Score(X) \geq S\} = 0$. This suggests sampling values of X that have higher scores. In doing so, we have to make sure that we are still computing their correct p-values with respect to P_{BG}.

Importance sampling (e.g. Gelman *et al.* (1995)) is a general method that estimates $\boldsymbol{E}_{P(X)}[f(X)]$ using samples from a *proposal distribution* $Q(X)$. The method relies on the following equality

$$\boldsymbol{E}_{P(X)}[f(X)] = \boldsymbol{E}_{Q(X)}[f(X)w(X)]$$

where the weight $w(x) = \frac{P(x)}{Q(x)}$ compensates for the bias introduced by sampling from Q.

To apply this scheme for estimating the p-values, we need to choose an effective proposal distribution Q. For the log-odds score, and most other scores, values of X sampled from the binding site model are more likely to receive high scores. Naively, we can set Q to be the distribution $P_M(X)$ of the binding site model, and directly sample from the region of high scores. This, however, is problematic since having low scoring samples is critical for the correct estimation of the p-values.

One possible solution is to sample a mixture of n_1 samples from P_M and n_2 samples from P_{BG}. While this solution takes into account both extremes, it still suffers from poor estimation of the “middle-ground” scores (results not shown). Thus, we refine the above approach and consider a richer combination of L models. We define *Compound Importance Sampling* (CIS) as:

$$Q(X) = \sum_{i=0}^{L} m_i Q_i(X)$$

where m_i is the fraction of samples generated from each model Q_i.

These models are basically “smoothed” versions of P_M that bias it in different degrees toward P_{BG}. In the case of PWMs, this is achieved by averaging the probabilities at each position between P_M and P_{BG}. In the general case, “smoothing” is done by averaging according to the dependency structure (for lack of space we omit further details). It is important to note, that this is not equivalent to mixing samples from the two extreme distributions. See Figure 1 (top) for an illustration of a proposal distribution. Figure 1 (bottom) illustrates a histogram of the p-value distribution from different components of the proposal distribution.

Finally, we have to decide on the number of components as well as the number of samples and degree of smoothing for each component. In this work we use 10 components and exponentially decay both the smoothing factor and number of samples starting from 10,000 samples for $P_{BG}(X)$ and 1000 samples for $P_M(X)$. Our CIS method proved robust to a large range of these settings (experiments not shown).

EXPERIMENTAL VALIDATION

As a case study, we examine the TRANSFAC 7.3 (Wingender *et al.*, 2001) model of RAP1 in *S. cerevisiae*, which is 14bp long. We generated 10^6 subsequences from a 3^{rd}-order Markov background model of *S. cerevisiae*. We computed the score of every site, and estimated its p-value using the following methods: CIS algorithm using 40,822 samples from a proposal distribution as illustrated in Figure 1; the MAST method (Bailey and Gribskov , 1998); functional approximation by Normal distribution, where we estimate the mean and variance of $Score(X)$ according to $P_{BG}(X)$, and then use the tail probability of Gaussian distribution as the p-value estimate. As a proxy to the truth, we computed the empirical p-values from

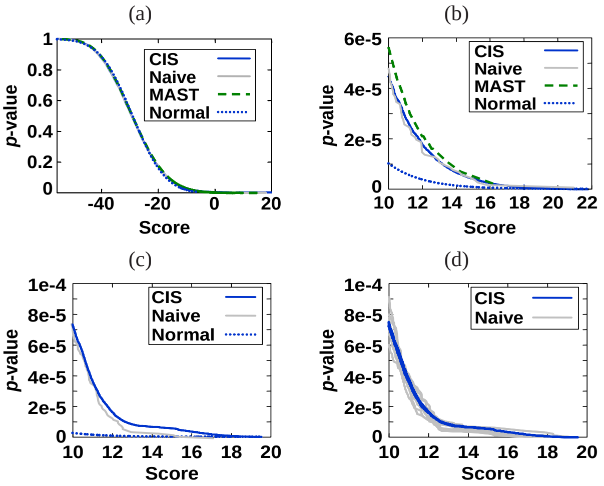

Fig. 2. Comparison of p-value estimations for binding site scores using Naive sampling (10^6 samples), the Normal approximation, MAST, and our CIS method (40,822 samples). The p-value (y-axis) is shown as a function of the log-odds score (x-axis). (a) For TRANSFAC position independent model of RAP1; (b) Same as (a), zoomed on p-values $< 10^{-4}$; (c) For the dependency model of PHO4 learned by Barash *et al.* (2003) (here MAST is not applicable); (d) Test for robustness with 10 repeats of (c).

the subsequences described above. Figure 2(a) compares the p-value estimates by the different methods. While all methods appear the same, zooming into the region of interest in Figure 2(b) reveals significant discrepancies. It is evident that the Normal approximation is inaccurate in this region. Both CIS and MAST provide accurate estimations, with a slight advantage of the CIS method.

Figure 2(c) shows evaluation of p-values for a binding site model with dependencies between positions learned by Barash *et al.* (2003) for the PHO4 transcription factor. For models such as this MAST is not applicable. As we can see, CIS's estimation are similar to the empirical ones. One might suspect that the slight deviation observed between the two is due to the smaller number of samples used by CIS. As Figure 2(d) shows, when comparing ten repetitions of the procedure, it is evident that CIS is more robust than the naive sampling, using two orders of magnitude less samples.

So far we showed the effectiveness of CIS with respect to the background distribution directly. We conclude by demonstrating our approach on a real-life genome-wide scan. We used the dependency binding site model of the ZAP1 transcription factor (Barash *et al.*, 2003) to scan the promoter regions of all genes in *S. cerevisiae* excluding those shown to be targets of ZAP1 by Lee *et al.* (2002). We thus expect that this set of promoters will contain few real binding sites of ZAP1. Again we use a 3^{rd}-order Markov model as a background distribution. Figure 3 shows that in this setting too, the Normal approximation results in poor p-value estimates. More importantly, the CIS method provides very accurate estimations over a wide range of p-values.

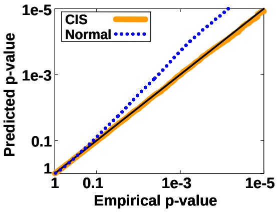

Fig. 3. Comparison of predicted and empirical p-values for a genome-wide scan of *S. cerevisiae* using the ZAP1 dependency model (Barash *et al.*, 2003), with a 3^{rd}-order Markov background model. p-value estimations using the Normal approximation and the CIS method are shown.

DISCUSSION

In this work we introduced a general and efficient method for estimating the statistical significance of putative binding sites in genome-wide scans. We demonstrated the accuracy of the method on both synthetic and genomic data, using simple as well as rich probabilistic models.

Given a motif model, another way to improve the statistical significance in a genome-wide scan is to use a more accurate background model. Since this issue holds crucial importance, we plan to explore methods for learning better background models in the future.

To our knowledge, this is the first method of its kind that can be applied to any probabilistic form of the binding site as well as background models. Its general framework makes it applicable for the identification of other sequence motifs, such as splice junctions, protein motifs, etc.

REFERENCES

Bailey,T.L. and Gribskov,Y. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.

Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling dependencies in Protein-DNA binding sites. *Proc. Seventh Inter. Conf. Res. in Comp. Mol. Bio. (RECOMB)*, 28–37.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J. Royal Statistical Society B*, **57**, 289–300.

Gelman,A., Carlin,J.B, Stern,H.S. and Rubin,D.B. (1995) *Bayesian Data Analysis*. Chapman & Hall, London.

Huang,H., Kao,M.J. and Zhou,X. *et al.* (2004) Determination of Local Statistical Significance of Patterns in Markov Sequences with Application to Promoter Element Identification. *JCB*, **to appear**.

King,O.D. and Roth,F.P (2003) A non-parametric model for transcription factor binding sites. *NAR*, **3119**, e116.

Lee,T.I. *et al.* (2003) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Sinha,S. and Tompa,M. (2000) A statistical method for finding transcription factor binding sites. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 344–354.

Wingender,E. et al. (2001) The TRANSFAC system on gene expression regulation. *NAR*, **29**, 281–283.

Wu,T.D., Nevill-Manning,C.G. and Brutlag,D.L. (2000) Fast probabilistic analysis of sequence function using scoring matrices. *Bioinformatics*, **16**, 233–244.

Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **to appear**.